



# Linking field synopses with data collection through disease-specific networks and biobanks (HuGE networks and P3G)

Julian Little and Isabel Fortier HuGENet Workshop, Atlanta, GA January 24-25, 2008



# How are we responding to the difficulty to identify and replicate genetic associations?

# Among others, by:

- Trying to conduct studies with optimal designs
- Increasing the size of individual studies
  - Very large sample size needed
- Promoting the conduct of meta-analysis
   Data pooling limited by the heterogeneity between studies (data access rules, designs, information collected, etc.)



# The Public Population Project in Genomics (P<sup>3</sup>G) choose to support:

- Sharing of knowledge: Support the development of biobanks and the exchange of expertise and tools.
- Documentation: Provide access to descriptive information on participant organizations (design, ethical rules, measures and samples collected, etc.).
- Harmonization: (1) Support the development of common guidelines; (2) Identify common information that could be shared across networks of studies
- Sharing of information: Support the development of IT tools facilitating data exchange.





# P<sup>3</sup>G Working groups:

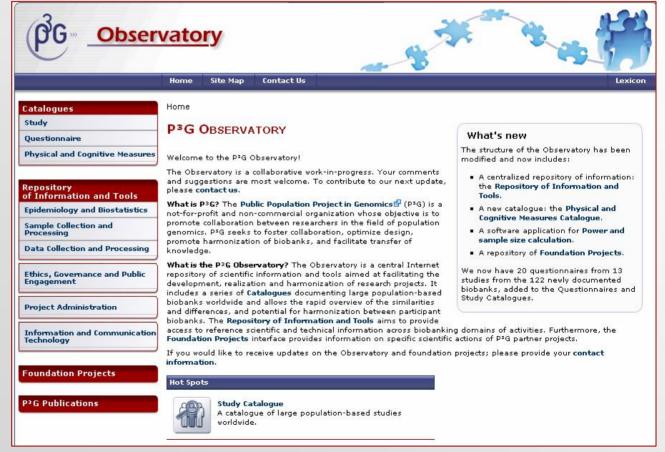
Genomics and Biochemical Investigations
Knowledge Curation and Information Technology
Ethics, Governance and Public Engagement
Epidemiology and Biostatistics



# The P<sup>3</sup>G Observatory: A collaborative work-in-progress

 Provide tools to support researchers in the harmonization, development, and realization of

studies.





# **Observatory Catalogues**

- They provide a quick and easy access to:
  - The information collected and methods in use by selected biobanks
  - Interactive tools allowing the rapid overview of the similarities and differences, and potential for harmonization between participant biobanks.

# **Study Catalogue**



## STUDY CATALOGUE

This Catalogue is a repository of standard information describing population-based studies in genomics.

### To be included in the Catalogue, a study must:

- Collect or plan to collect DNA samples or biological material from which DNA can be extracted.
- Have recruited or plan to recruit more than 10 000 healthy individuals (some smaller studies are also included occasionally).



The descriptive information is collected in a two-steps process: (1) The summary information is initially gathered by our team through websites and/or publications and (2) Complete information is provided by study investigators using the P<sup>3</sup>G study description form. For studies with only a summary description, the scope and quality of information collected is limited, and the search form provides fewer data. If you would like your study to be included in the catalogue, please complete the **study description form** and send it back to us by **email**.

Catalogue Current Content		
Number of Studies	122	
With Summary Information Only	81	
With Complete Information	41	

### Comparison Tools

Three tables are provided in xls format to compare information on studies.

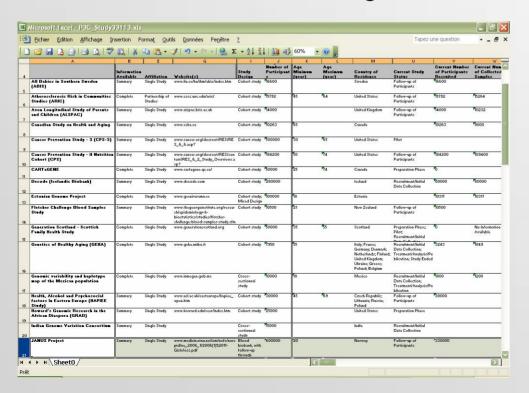
Table	Description
Summary Information	Summary information.
Principal Topics of Interest	Principal information collected.
Ethics & Governance	Information on ethics and governance.

122 large population-based biobanks

# **Study Catalogue**

Description of the population-based biobanks (contact information, website, objectives, design, selection criteria, current status, etc.)

**Comparison Tables** (Summary information; principal information collected; ethics and governance issues)









# Number of participants targeted (recruited or to be recruited) (N=105)

Number of participants	Number of studies	Number of participants TARGETED
Less than 49 000	60	1,100,000
50 000 to 99 000	15	1,000,000
100 000 to 499 000	23	3,600,000
500 000 and more	8	5,600,000
		Total: 11,300,000



# Coming soon...

- A catalogue of networks will be added.
  - Studies part of a network will be described using the standard description format
  - Will include different types of networks (diseaseoriented networks could be included)
  - Will be develop in collaboration and will use a common interface



# **Questionnaire Catalogue**

## **CROSS-SECTIONAL QUESTIONNAIRE CATALOGUE**

The Questionnaire Catalogue includes reference questionnaires from large population-based studies and aims to facilitate the development and harmonization of questionnaires. The catalogue gives access to the cross-sectional questionnaires, their methods of administration, and provides an overview of the similarities and differences of the information collected by participant studies. References to other questionnaires of interest are also available in the Repository of Reference Questionnaires.

Through a tree of key-words, a **Search Tool** gives quick and easy access to the information collected by a given study and allows comparison between studies. A **Comparison Table** of the information collected by the studies can be viewed via the web browser or can be downloaded as a table in **Excel format**.



### **Catalogue Current Content**

Number of Studies 13 Number of Questionnaires 20

### List of Questionnaires

1998 Health and Social Survey (HSS)

1998 Health and Social Survey; Lifestyle and Health Questionna

13 studies; 20 questionnaires

Cancer Prevention Study - 3 (CPS-3)

CPS Questionnaire for Men

**CPS Questionnaire for Women** 

3. Estonian Genome Project (EGP)

**Estonian Genome Project Questionnaire** 

4. European Prospective Investigation into Cancer and Nutrition (EPIC)

24 Hour Diet Recall

# **Questionnaire** Catalogue

# General information, methods and questionnaires

### QUESTIONNAIRE

### Question Block Annotation Search Questionnaire

### General Information

Name Estonian Genome Project Questionnaire Study Estonian Genome Project (EGP) EGP-O-1.00; EGP-O-2.00; EGP-O-3.00 Version

Estonian Genome Project of University of Tartu Author

Restriction On Utilization

Contact

Prof. Andres Metspalu Estonian Genome Project EGP

Tiigi 61B

50410 Tartu Estonia Phone: +372 7 440420

Email: Andres.metspalu@geenivaramu.ee

### Information On Validity

Validation of the questionnaire has not been carried out yet

### References

No Information Available

### Methods

### Respondent



✓ Participant

X Proxy

### **Administration Environment**

- X Over the phone
- Hospital, Clinic, University or Recruitment Center
- X Respondent / Proxy Residence

### **Administration Mode**

- X Auto Administered
- X Auto Administrated With Face To Face Validation By Trained Personnel
- ✓ Administrated By Trained Personel / Physician

### Administration Format

- X Paper Questionnaire
- ✓ Computerized

Administration Language Estonian

### Documents

### **Available Format**





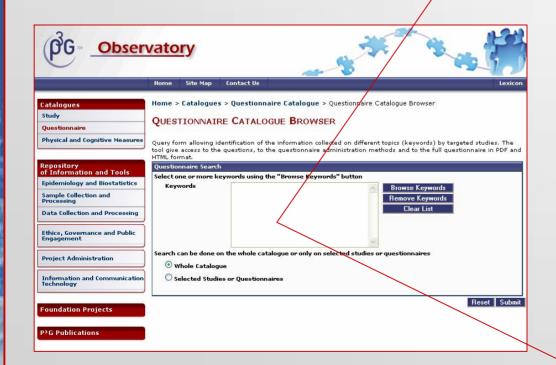
NOTICE:

HTML document format may differ slightly from the actual questionnaire.

Please refer to the original PDF document for exact reproduction.



# **Questionnaire content: Search by keywords**



🚵 Health Information Health assessment Medical care system/procedures ★ Medication intake (ATC/ADD) ⊕ 🚵 Women's health H Men's health Perception of health/quality of life Physical Conditions Anthropometric information > Handedness Life Habits/Behaviours >> Smoking/Tobacco use Alcohol intake Drug intake 🗓 🧐 Physical activity Weight related behaviours Sleep Sexual behaviours Travels: Sociodemographic Characteristics Birth location Subjects birth location Subjects family birth location Citizenship/residency status >Ethnicity/Race Language Marital status Education level 🕀 🔧 Working status Religion Physical Environment 🖭 🔞 Environmental exposures SEarly life and in utero exposures

Social Environment

SFamilial and social environment
SWorking social environment



# **Comparison table**

Life Habits/Behaviours	BNADN	STR	HSS	KORA- gen	MCPMR	KSCDC	ТТР	NHANES III	EGP	EPIC	CPS-	GS- SFHS
Smoking/tobacco use	1	1	1	1	1	1	1	1	1	1	1	1
Alcohol use	1	1	1	1	1	1	1		1	1	1	1
Illicit drug use		1	1						1			
Nutrition	1	1	1	1		1	1	1	1	1	1	1
Food intake list	1	1		1		1	1	1	1	1	1	4
Intake of milk products	1	1		1		1	1	4	1	1	1	*
Intake of meat, eggs, fish and alternatives	1	1		1		1	1		1	1	1	1
Intake of vegetables and fruits	1	1		1		4	4	4	1	1	1	4
Intake of cereals, bread, starches	1	1		1		1	*	-	1	1	1	1
Intake of sweets, baked goods	1	4		1		8	*		1	*	1	
Intake of fat		1		1		38	1			1	1	
Intake of water				1		s	1				1	
Intake of beverages (other than water)	1	4		1		1	*		1	4	*	
Intake of salt and food supplements		1				1	1	1	1	1	1	1
Perception of nutritional habits		4	1				8 9	7		1		



# **Questionnaire catalogue: Bloc of questions**

eneral Information				
Block ID	100073_11			
Study	National DNA Bank - BancoADN (BNADN)			
Questionnaire Name	Spanish DNA Bank Questionnaire			
Keywords • Alcoholuse				
uestion(s)				
	est efforts, the formating of the text below may differ slightly from the actual questionnaire. Please recument for exact reproduction.			
ALCOHOL				
32. How much alcohol do	you habitually consume?			
Wine:				
Do you drink wine with :				
Yes, how man	y glasses (10cl)?			
Do you drink wine outsi	de mealtimes?			
Yes, how man	y glasses (10cl)?			
No				
Beer:				
Do you drink beer every	day?			
Yes, how man	y beers (33cl)?			
_    No				
Do you drink beer at the				
Yes, now man	y beers (33cl)?			
11 NO				
Spirits:				
Do you drink spirits ev	ery day?			



# Physical and Cognitive Measures Catalogue

## PHYSICAL AND COGNITIVE MEASURES CATALOGUE

IMPORTANT! This catalogue is currently under piloting. For any comments or suggestions, please contact us .....

The catalogue provides an overview of the physical and Comparison of the participant biobanks. When available, it also possent to some and operating out of the Comparison Table allows comparison of the similarities and differences of the measure collected among studies through a tree of key-words, developed upon the WHO International Classification of Functioning, Disability and Health (ICF).



### Catalogue Current Content

Number of Studies

- 6

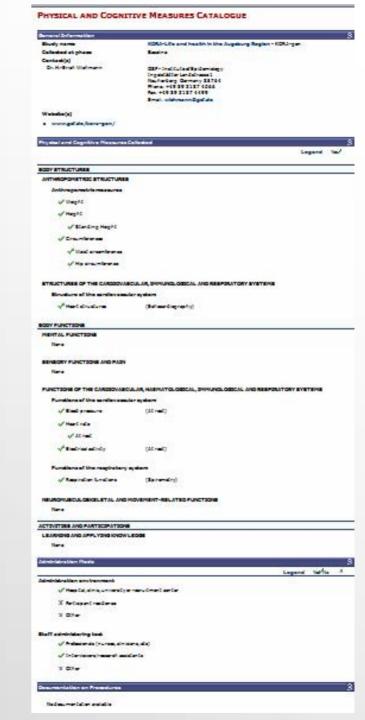
### List of Participating Studies

6 studies

- 1. KORA-Life and health in the Augsburg Region (KORA-gen)
- 2. Kadoorie Study of Chronic Disease in China (KSCDC)
- 3. Marshfield Clinic Personalized Medicine Research Project (MCPMR)
- 4. National DNA Bank Banco ADN (BNADN)
- 5. National Health and Nutrition Examination Survey III (NHANES III)
- 6. The Tomorrow Project (TTP)

# Physical and Cognitive Measures Catalogue

- General information
- List of measures collected
  - Body structures
  - Body functions: based on the International Classification of Functioning, Disability and Health (ICF-WHO)
- Administration mode:
- Standard operating procedures (if available)





# PHYSICAL AND COGNITIVE MEASURES KEYWORDS

Messures keywords	Studies					
BODY STRUCTURES	BNADN	KSCDC	KORA- gen	MCPMR	NHANES	TTI
ANTHROPOMETRICSTRUCTURES	1	1	1	*	1	4
Anthropometric measures	1	1	1	1	1	4
Weight	1	1	1	1	1	4
Standing height	1	1	*	1	1	4
Sitting height		4			1	
Waist circumference		1	1		1	*
Hip circumference		4	1			
Other anthropometric measures	0	3.5c			4	
Body composition		1		4	1	
Bioimpedance		1			1	
Bone density				1	1	
Skin fold thickness		80			1	
Other body composition measures					4	
STRUCTURES OF THE CARDIOVASCULAR, IMMUNOLOGICAL AND RESPIRATORY SYSTEMS*		(a)	1	e e		

BODY FUNCTIONS*	BNADN	KSCDC	KORA- gen	MCPMR	NHANES	Т
MENTAL FUNCTIONS*		1		1		
Global mental functions (b110-b139)*		1		4		
Intelligence						
Congitives functions		1		1		
Other global mental function measures		,				
Specific mental functions (b140-b189)*					1	
Attention	15	× = = = = = = = = = = = = = = = = = = =				
Memory		17				L
Handedness					1	
Psychomotor function		ev-			4	
Other specific mental function measures						
SENSORY FUNCTIONS AND PAIN*		5		4	4	
Seeing and related functions (b210-b229)*				1	1	
Vision				1	4	
Hearing and vestibular functions (b230-b249)*					1	
Hearing					1	
FUNCTIONS OF THE CARDIOVASCULAR, HAEMATOLOGICAL, IMMUNOLOGICAL AND RESPIRATORY SYSTEMS*	1	1	1	1	1	
Functions of the cardiovascular system (b410-b429)*	1	1	1	4	4	
Blood pressure	4	1	1	1	1	
Heart rate (at rest)	1	1	*	4		
Heart rate (under stress)						Г



# **Next steps:**

Next catalogues update, May 2008

**Studies** 

**Networks** 

Questionnaires

Physical and cognitive measures

**Ethics and governance** 

**DNA** processing and storage

 Future developments: Addition of new catalogues and inclusion of the longitudinal information collected





Provides access to reference information (documents, websites, etc.) in different domains of biobanking:

Repository of Information and Tools

**Epidemiology and Biostatistics** 

Sample Collection and Processing

**Data Collection and Processing** 

Ethics, Governance and Public Engagement

**Project Administration** 

Information and Communication Technology



# **Repository of Information and Tools**

# Structure: Bioscience section

Conceptualization
Design and Conduct
Analysis
Dissemination

In Theory (guidelines...)
In Practice (SOPs...)
IT Tools

# **Epidemiology and Biostatistics Examples:**

The HuGENet™ HuGE Review Handbook, version 1.0

# Conceptualization: In Theory: HuGE Review Handbook

Conceptualization:
IT Tools:
ESPRESSO Power
Calculator

Observatory

Physical and Cognitive Measure

**Epidemiology and Biostatistics** 

Data Collection and Processing

Information and Communication Technology ESPRESSO POW

program for simulation-base

Population prevalence of disease

Population prevalence of 'at risk' genotype

Odds ratio (OR) associated with 'at risk' genotype

Population prevalence of 'at risk' level of environmental factor

John Gainacher, Maria Gwinn, Junan Friggins, John Rommon, John Knoury, Sarah Lewis, Julian Little, Teri Manolio, David Melzer, Cosetta Minelli, Paul Pharoah, Georgia Salani, Simon Sanderson, Lian Smeeth, Lealey Smith, Jonathon Sterne, Donan Stroup, Emanuela Taioli, John Thompson, Simon Thompson, Neil Walker, Ron Zimmenn)

nd HuGE Net Executive Group (Molly Bray, Marta Gwinn, Julian Higgins, Jol cannidis, Muin Khoury, Julian Little, Teri Manolio, Ron Zimmern)

USDA/ARS Children's Nutrition Research Center, Baylor College of Content, Northead Usak, "MCR Biotection Usak," and Biotection Usak, "Intelligent States Usak," and Biotection Usak, "Intelligent States Usak," and Biotection Content of Biotection Content States Usak, "Intelligent States Usak, "Clinical evidential Epidemiology Usin, bepartment of Hygiene and Epidemiology Content of Commins School of Medicine, Biotection of Research and Technology—Hells, commins 43110, Geneses 10 er of Genomics and Disease Prevention, CDC, Athanta, Oxida, Usak, "Commission of Commission Content of Content Content of Content Content of Content Content of Content Conte

PAGE ACTIONS

HISTORY

GENESTAT

GeneStat provides a knowledge base for statistical genetics through an Internet-based series of tutorials and reviews with links to key sites and computer programs for analysis of genetic data.

Endown the company of the company of

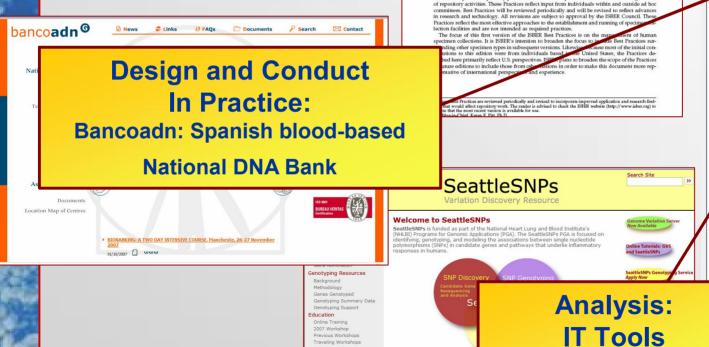




Design and Conduct, In Theory:

Best Practices for Repositories I:
Collection, Storage, and Retrieval of
Human Biological Materials for Research (ISBER)

Investigator Opportunities
SeattleSNPs offers investigators



EPIDEMIOLOGY AND BIOSTATISTICS

In Theory

**SeattleSNPs** 

The application of the principal of the control of

C 1001 Adia Paul de Pend o Germa

Data collection and processing **Examples:** 



**Design and Conduct:** In Theory: **Classification indexes** (occupation, diseases, mental disorders, body functions, etc.)

STATISTICS

**About National Statistics & ONS** 

nite papers" have been



14 November 2006

for Children's Environmental Health and Disease Prevention Research

Ethics, Governance and Public Engagement



Legislations specific to biobanks
Selected Guidelines
Selected opinion, background documents, etc.

ina 144

Law Roform Commission (Australia, 2003)

Haddow and Cunningham Burley, Generation Scotland: The Scottish Family Health Study Public Consultation: Discussion Paper. 2005, #[Nic consultation research project for Generation Section (Section)].

In Practice:

Application of theory/legislation to a given biobank

ent of Principles on the Ethical Conduct of Human Genetic Research Involving Populations This Microcol.

spics on the Ethical Conduct of Human Genetic Research Involving Populations is besed on a framework of
famonial principles giving rise to specific recommendations and procedures for their implementation. Network
and Genetic Medicine (Canada, 2003).

opinions, background documents, literature, and other ELSI material

lce

n of the theory/legislation to a given biobank. What happens in reality? How is the law translated in practice? appert on public consultation conducted by a population-based biobank; consent forms or practice are adopted by a given biobank.

tion Scotland - Legal and Ethical Aspects Toppoper provides an overview of the Issues to strike an

riate beliance between the public interest in ensuring that the value of Generation Section is realised and brought to the Section people and, second, the interest in ensuring that individuals who perticipate in the project are

Tools:

ers of the public and

d Stakeholders Views

PopGen, HUMGEN

Web search tools and PFG template

HUMGENON: website includes various research tools and presents social, othical and legal aspects of human genetics. Issues such as confidentiality of genetic data, consent to genetic testing, and stem cell research are explored.



# **Project Administration**

Home Site Map Contact Us Lexicon

Study Questionneire Physical and Cognitive Measures

Home > Repository of Information and Tools > Project Administration

### PROJECT ADMINISTRATION

This scotion provides references and tools to support administrative activities of population-based studies in genomics. Please visit the following sections:

- . Ethics, Governance and Engagement, for governance considerations:
- Sample Collection and Processing, for laboratory infrastructures.

Epidemiology and Biostatistics

In Theory

Sample Collec Processing

Date Collection

Ethics, Gover

Project Admi

- In Practice IT Tools

Information and Communication Technology

In Theory:

Best practices, guidelines for project administration

Remacives. Published by the Organisation en Siological Materials for Research

nt comprises the report on best practice

ground and rationals to the project and

d provides general recommendations on

management.

of repository activities on the reflect the most effective approaches to ntended as required practices. They will

be reviewed periodically and will be revised to reflect advances in research and technology. Prepared by the International Society for Biological and Environmental Repositories (ISSER) (2003).

### In Practice

Scal-life examples and information related to administration of socials projects.

The National Children's Study: Detailed Summaries of Sample Size and Budget Constrained Cost Estimates ndie H) #Tument helps establishing a study start up budget. "White paper" developed for The National ly, Masupported by U.S. Environmental Protection Agency, U.S. Department of Health and Human

ildren's Study: Detailed Summaries of Sample Size and Budget Constrained Cost Estimates and ions Using Revised Retention Retes (Appendix I) doffinent helps establishing a study start-up evised retention rates. "White paper" developed for The National Children's Study, Cost@ported by ntal Protection Agency, U.S. Department of Health and Human Services, USA.gov.

# In Practice:

**Examples of budget** estimations

for project administration.

Taskjuggler 🗹 an open source project management software for serious project managers. It covers the complete spectrum of project management tasks from the first idea to the completion of the project. It assists you during project scoping, resp

Source software is KPleto (7the Koffle

being developed for

Open Workbench it and management fi CA's Clarity Division

dotProject toan op manage tasks, scho supported by a vol-

loroject-open to a arcas such as CRM.

PresMindt a free substacks and time

## IT Tools:

Selected software to support project administration

CPlate is

duling veloped by

a teal to

ntegrates icinita.

state fo

# Information and Communication **Technology**



a project, please write to info@obiba.org

### Obiba

- Overview > Vision
- > Why open source?
- > P3G Core
- > Online Survey
- Software Directory
- > Partners

### Projects

> GenoByte

> Sample Mana

### Overview

Obiba is an open source project whose aim is to build an open software software infrastructure on which biobanks around the world can depend and to which they can contribute. To know more, see our

Ohiha nartners are currently working on the two areas of

development which are listed in the left column under Projects. All these projects welcome collaborators. If you are interested in getting

involved, please join our mailing lists. To ask specific questions about



Obiba Funded



# **Example:**

**Obiba Website** Open source management system for biobanks

### INFORMATION AND COMMUNICATION TECHNOLOGY

This section lists the software referred to in the other sections of the Repository.

	-	Significan		1
	Epidemiology and biostetitics	Sample collection and processing	Date collection and processing	Project Administration
American National Center for Siotechnology Information (NCSI)		1		
Bioconductor	1	1	(c) (c)	KG 24
Broad Institute Software Including Arachne, Argo, Conrad, PLINK, EIGENSTRAT, Haploview, Locusview, Sweep, Togger, Geneflunter, MapMaker3, GenePattern, GSEA, GeneCruiser, ConnectivityMap	-			
CLC Free Workbench		1		
dotProject	3 3		22 23	1
Dr Devid Cleyton's website Including SPLINK, TRANSMIT, PEDZSPL, GHZSTAT and SNPHAP	1	,		
Dr Devid Reich's Leboretory Including ANCESTRYMAP, EIGENSOFT and EIGENSTRAT.		4		
Epi Info	4		1	77
ESPRESSO Power Calculator	1		0 0	
European Signiformatics Institute (ESI)		1		
PresMind	2 3		3 - 3	1
GeneSNPs		1		
GeneStot	1	4	× ×	
Genetic Association Database (GAD)		J		
Genetic Power Celculator		4		
Gonçalo's Software	4	1		
HepMep Including genome browser and the Encode Project.		1		
KPleto			9 - 0	-
Matthews Stephens's website noluding SIMSAM, fastPHASE, HOTSPOTTER, PLASE and SCAT		~		
The Obibe Project		<b>√</b>	ya sy	62 5
Open Workbench				1
PEDSYS	4	1	1	-1



# **Foundation Projects**

## **IWG1**

- **DNA Quantity and Quality** Control (Q2C)
- Harmonization and Quality Control of Lipidomics Analysis

## IWG2

The Obiba Project

## IWG3

P3G Policymaking Core

## **IWG4**

- Generic Dataset
- GeneStat
- **ESPRESSO** Power Calculator

# P<sup>3</sup>G Observatory



Harmer Promisition Projects

### FOUNDATION PROJECTS

Transported person interesting or PC to delive project, and transect, install aproject extent followed are the deliver project, aproject has below PC seven to a project operations in the other project, a project has been PC seven to a project of perfective interest that is part of the PC in Enthropie Colorest with a Challen. They may it this between the matter or all they are a related to the lat.

International Working Group 1: Genomics and Blochemical Investigations

### DNA Quantity and Quality Control (Q2C) 4

Main stally avaisis Informational guidelines for CRA measurements.

International project Coherence (INI) measurements in each mission and service of the International Coherence (International Coherence) of the International Coherence (Initernational Coherence) of the International Coherence (Initernational Coherence) of the Initernational Coherence (Initernational Coherence

Correction and Correct

Sale Collection and Press

Biblio, Governmence and Public

Harmonization and Quality Control of Lipidomics Analysis Contact parent

Majo sindo modele Continued in the Continued of Assertation in Indiana asserts

The project area failed by harmonical and finish communications of related entition of them a not a started ordinate and qualitative for the forms

International Working Group 2: Knowledge Curation and Information Technology

### The Obliba Protect 5

Contact recent Vingant Percentit Co.

Main sinformation Underlanded by an integrated activation accomplishment in management

The CO Co Project will provide an upon source information management update for Contaction (And see Contaction in the Constant and This information system will

support a side range of the tertilative and self-ratio to the tertilar per interferent a tertilation from the term of the self-ratio terms.

### International Working Group 3: Ethics, Governance and Public Engagement

PFG Policymaking Core

Contact parent Bartina Plania Kompysin C.

Earne d'agres et le commité de l'article par à communication, d'als assesses de la communication, et à Communication de proposition d'administration de

The project size to below retain their formations the course and submedien is

palay approach as and madificating in appoint Conformer grand is an identiting to

### International Working Group 4: Epidemiology and Biostatistics

Contact person

Sealers Further La

before the first installing a left of problem of profession in formal to be used by and

should believe energing bisharis.

by large age if which great also disc and disclorate wherever the fundamental size is

armidea form infacts built of a former indice but one binder is and a paint the

GeneStat \*

Under Lawrenchine, a name of Laborate and records for any one of control data. Condition are not an element of a large transfer also falled and person in the supply on the forest-

american la maiori d'amolio dalla

### ESPRESSO Power Calculator

Notice and and confine places are referring over and completion

SEMBERS (Relimating Respirational Four in Rily Expiring Resolded Blody) Outcome) is a new Wildowski program for simulation-based power saturation. That are to used to estimate resisting complication may be seen for the constant for erd artest alvalar in projection generalis information become after direct

pediar extremely distant clouders.

### PTG Observatory

### Catalogues

Contact recent

Terim of and inCortem installing the Blacky Quantitarymains, Physical and Main sindy available

Cognitive Measures, 90ths and German search Lampis Europe and New Colony view.

The Calleguine provides a published any assess to information on providing to propulation - Count Coulombs and also arrapid morning of The similar Countries Offerense, and automballe harmonisation between participating biobants

### Theory and Practices Repository

Sector Parties and Contact persons Majo dally avaisin

that interference thing assess for elementaries with the indifferent



# One example: The Reference Datasets

# Aim:

 Provide a template to facilitate harmonization between biobanks and support the design of emerging ones.

BB 1

BB 2

COMMON

BB 3

BB 4

# **Characteristics of the Generic - Reference Dataset:**



- Set of variables that is comprehensive enough to ensure the realization of valid research
- Simple enough to allow implementation in a variety contexts
- Small enough to encourage buy-in.
  - NOT a prescriptive list of all the variables to be collected by a biobank. Approximately 250 variables.
- The first initiative focused on middle-aged subjects and information collected at baseline
- Complementary to development of specialized datasets for particular interests (e.g. particular diseases, environmental exposure, etc.).



# Some examples of domains covered

## **Health outcomes**

 Cancer; diabetes; stroke; myocardial infarction; familial history of cancer, etc.;

## **Health determinants**

 Smoking, alcohol intake, birth location (subjects, parents and grand-parents), education, income, passive smoking exposure, working status, physical activity

# Physical measures

Anthropometric measures, resting heart rate, blood pressure

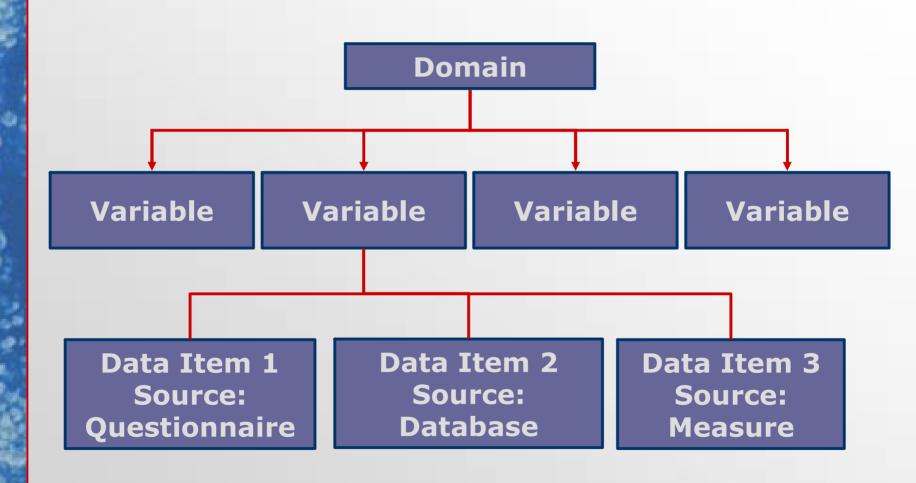


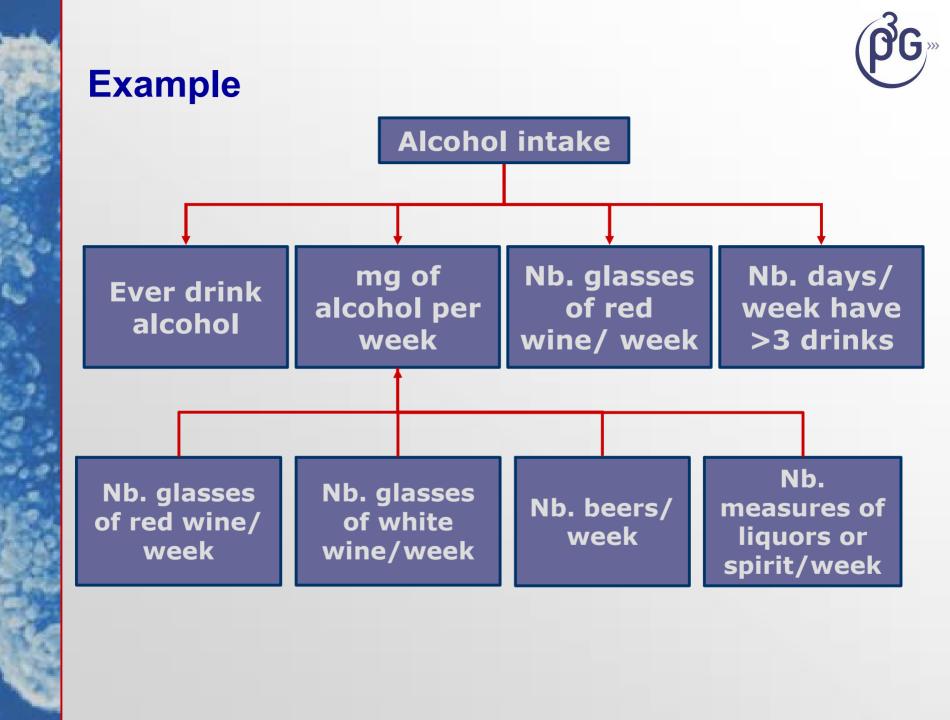
# **Current status:**

- Initial version (January 2007)
- Piloted by 4 biobanks
  - Lifelines (Netherlands), Joondalup Family Health Study (Australia), CARTaGENE (Canada), Cardiovascular survey (Canada)
- Updated version to be finalized: March 2008
- Web interface under development: Online May 2008
- New reference datasets (Cancer, etc.) to be developed



# **Conceptual Framework**







# Coming soon...

- Online interactive tool available
- The web interface will provide access to:
  - Technical information:
    - Definitions and formats
    - Link to relevant ontologies
  - Scientific information:
    - Relevance
    - References
    - Coverage among participant biobanks (%)
    - Link to potential questions (or SOP)
    - Etc.
- Could host other reference datasets (diseasespecific, etc.)



# **Domain**

Domain

Risk fator or outcome of interest. A domain would include

Generic dataset
Datasets
Themes
Domains
Variables
Data items

**Publication** 

Admin

GENERAL INFORMATION Individual Cancer History Individual Disease History Theones

Definition A record of a patient's medical background regarding the occurrence of cancer and cancer-religied problems Date of assessment of coverage In total, 12 of 13 studies (92%) in the P30 Questionnaire Catalogue recorded

information (at recruitment) about the domain of interest.

(Based on Guestonnaire Catalogue's Reyword : Neoplasms) Inchesion status true. Catalogues Questionnaires

Under development

Classification level Scientific information being collected

ONTOLOGY / CLASSIFICATION

Completion status

Outology name NOT Then away VOID Keyword name Pichyldusi Concer History: Neoplasms C18848 (C00-048)

Reference Cancer is of great public heath relevance, it is a leading cause of death across the world and causes substantial mortality. Of the 5 deaths worldwide in 2005, cancer accounted for 7.6 million (or 13%) (1). In industrialized nations more than 25% of people will de 1 cancer (2) and the number of these deaths is projected to keep rising across the world. It is estimated that 9 million people will de-tricancer in 2015 and 11.4 million in 2030 (1). Cancer may affect people at all ages, but for most common cancers the risk increases in with age. Nearly two thirds of cases are diagnosed in people aged 65 and over (3). Many genes and the stylelenvironmental exposiimportant determinants of cancer. Many of these determinants are potentially subject to intervention and are therefore, well worthly scentric study within a biobank. Cancer presents a prevalence that allows adequate statistical power. Evaluation by questionnaire reliable measures. There is therefore a strong case for collecting information about concer per se in a bootank and also for ensuring information about cancer status is collected at recruitment into a cohort biobank. Associations Cancer is caused by both enternal factors (flobacco use, environmental tobacco use, physical fractivity, obesity, poor det, exposur ultraviolet radiation, exposure to certain viruses and bacteria, certain horisones, certain chemicals and other substances, ionizing re and alcohol-) and intrinsic factors (inherted mutations (heredity), hormones, immune conditions, and mutations that occur from metal and age (5, 7). These causal factors may act together or in sequence to inflate or provide cardinogenesis. Ten or more years often between exposure to external factors and detectable cancer (5). Comments

Scientific 1) Cancer 2007 World Health Organisation. http://www.who.int/mediacentre/fractisheets/frs297/en/index.html references

2) World Cancer Report, 2003, World Health Crosnization. 3) Cancer Research UK (Jan 2007): UK cancer incidence statistics by age. http://info.cancerresearchuk.org/cancerstatis/incidence/lager

4) National Cancer Institute, 2006. Facing Forward; Life After Cancer Treatment, 60 p. NBH Publication No. 06-2424. Http://www.cancer.gov/cancertopics&re-after-breatment.pdf

5) American Cancer Society. 2007. Cancer Facts & Figures 2007. http://www.cancer.org/downloads/STT/CAFF/200TPV/Secured.pdf 6) Canadian Cancer Society. 2006. Environmental risk factors for cancer.

http://cancer.ca/tocs/internet/standar.pt/j.2102.3172\_1435019236\_jargid-en.00.html T) University of Texas M.D. Anderson Cencer Center, 2007, Causes of cencer

http://www.mdanderson.org/batients\_public/about\_cancer/display.ctm?id=411A2531-7789-11D4-AEC3005088CCE3A8methodi-displayFul

Associated

General justification and limits of the	Variables include the occurrence of cancer as well as the type and onset of the cancers
variables	diagnosed. These variables have been selected by a committee of experts and based on the compensors of the baseline questionnanes of 15 population-based biologists. There was general agreement for the inclusion of those information and the reasonable validity of the information generated.
Specific references	-
Full block of questions and procedures	Y W30-obs_communit - GENERIC DATASET@sock of
	SWINDOWS CANCER DOMAIN GUESTIONS BLOCK 104-12-2007 (80)
	* It is also possible to present the current section under a table format
	1) Has a doctor ever told you that you had cencer?
	Instructions: None
	0= No
	1= Ves
	8= Freter not to enawer
	S+ Conft know
	Skip pattern: if Tito" or "Prefer not to answer" or "Don't know", go to Next domain

**Variables** 

Ask only HEVER HAD CANCER + 1 (Ves)

# **Data Items**

Title	tie EVER_HAD_CANCER				
Variable	9	Occurrence of cencer			
Domain SHERAL INFORMATION SOURCE					
Definition	Title	EVER_HAD_CANCER			
	Variables - Occurrence of concer				
	Pomains + Individual Cancer History				
	Source of information	Questionnaire			
i	Definition	Question asking the participant if he ever had	a cancer diagnosed by a doctor.		
	Question or SOP	Has a doctor ever fold you that you had cance	er?		

Title	EVER_HAD_CANCER
Variables	Occurrence of cencer
Domains	Individual Cancer History
Source of information	Guestionnaire
Definition	Guestion asking the participant if he ever had a cancer diagnosed by a doctor
Question or SOP	Has a doctor ever told you that you had cancer?
Categories	0=No 1=Yes 9= Don't Inow
Format	Numeric (categorical)
Criteria for transfer to the data item	

Criteria for

Definition

_			Occurrence of cancer     Individual Cancer History	
- 1				
EHERAL	INFORMATION			4
Title EVER_H		EVER_H	IAD_CANCER	-
/ariable	ariables - Occu		erence of concer	-
omains	SENERAL INFORMATION		1	
efinitio			EVER_HAD_CANCER	
Duestion	Variables		Occurrence of cencer	
Categori	Domains		Individual Cancer History	
	Source of information		Organización	

Has a doctor ever told you that you had cancer?

Question asking the participant if he ever had a cancer diagnosed by a doctor.

Pormat Catagorian	Cable		
SEHERAL INFORMATION			
Title	EVER_HAD_CANCER		
Variables	Occurrence of cencer		
Domains	Individual Cencer History		
Source of information	Guestionnaire		
Definition	Guestion asking the participant if he ever had a cancer diagnosed by a doctor.		
Question or SOP	Has a doctor ever told you that you had cancer?		

Categorgeneral Information	
Title	EVER_HAD_CANCER
Variables	Occurrence of cancer
Format Domains	Individual Cancer History
Criteria Source of information	Questionnaire
CommcDefinition	Guestion asking the participant if he ever had a cancer diagnosed by a doctor.
Criteria Question or SOP	Has a doctor ever told you that you had cancer?
Categories	0+No 1×Yes 8= Peter not to answer 9= Don't Innov
Format	Numeric (categorical)
Criteria for transfer to the data item	The second secon
Comments	
Criteria for transfer to next data item	If 0(No) or 8(Prefer not to answer) or 9(Don't know), go to Next domain